Management of (Ocean) Observations

- a report from OceanObs09 -

Steve Hankin, NOAA/PMEL

GO-ESSP

Oct. 2009, Hamburg, Germany

OceanObs'09

Ocean information for society:
sustaining the benefits,
realizing the potential

21-25 September 2009, Venice, Italy

Follow-on from OceanObs99

- 600 participants
- *in situ* / remote-sensed
- global / coastal
- marine biology, fisheries, ecosystems
- data assimilation

"access to obs"

A phrase we often use in planning
model integration …
but where is the coherent plan?

(Expectations for obs **integration**
exceed investments.)

DM on ocean obs is a hard problem

- multiple (in)distinct data structures

- granularity is an issue (esp. metadata)

- differing requirements and priorities separate RT, archive and "operational"

- many uncoordinated lines of funding

- legacy formats, traditions

The "social" problem is even harder
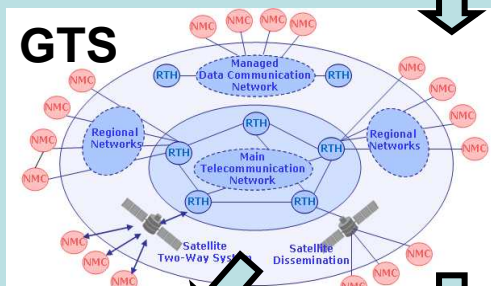
Competing (diverging?) standards:
- WMO / JCOMM
- IODE
- OGC – WCS, WFS, SOS
                    (several approaches to each)
- CF-netCDF-DAP
- DAPPER
- ERDDAP

Obs

Schematic
- global ocean obs -

GTS

Telemetry

Data centers

assimilation centers

archive centers

data assembly centers
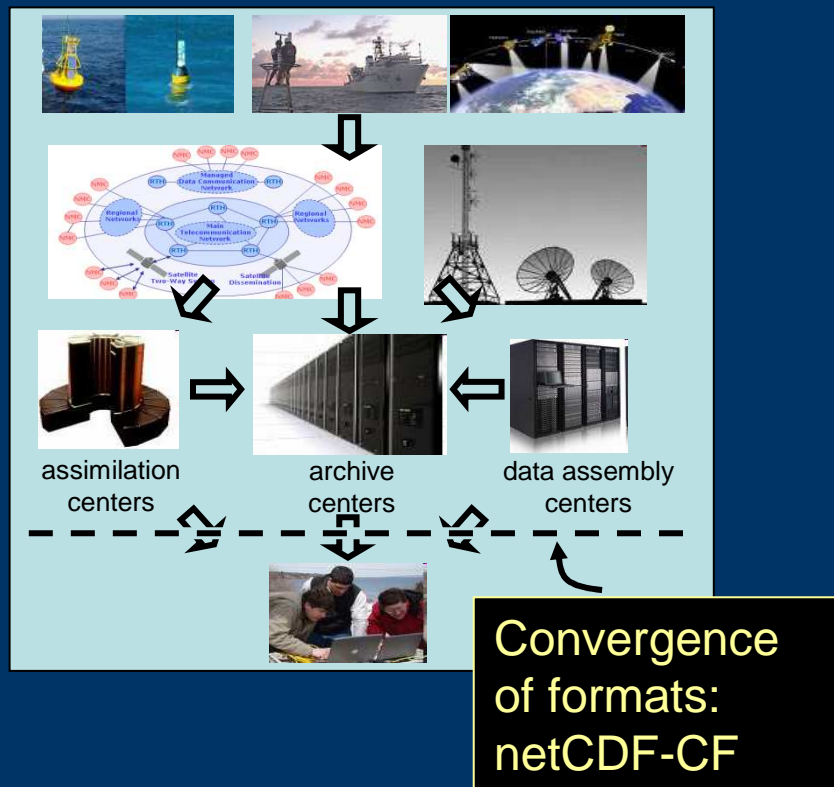
Convergence on uniform file formats: netCDF-CF(*)

*OGC standardization process underway 9/2009

OceanObs09: build on this momentum: create and share netCDF-CF files

Convergence of formats: netCDF-CF

assimilation centers   archive centers   data assembly centers

I. Complete:
   Argo, OceanSites, AVISO, GHRSST, …

II. Add:
   underway obs, tide gauges, other satellites, Ocean Atlas, GTSPP, CPR,…

III. Install THREDDS and OPeNDAP servers. Aggregate.

10-100 million files

Ocean obs

In situ   satellite

ArgoGDAC OceanSites   TAO globalDrifterCenter WODCs
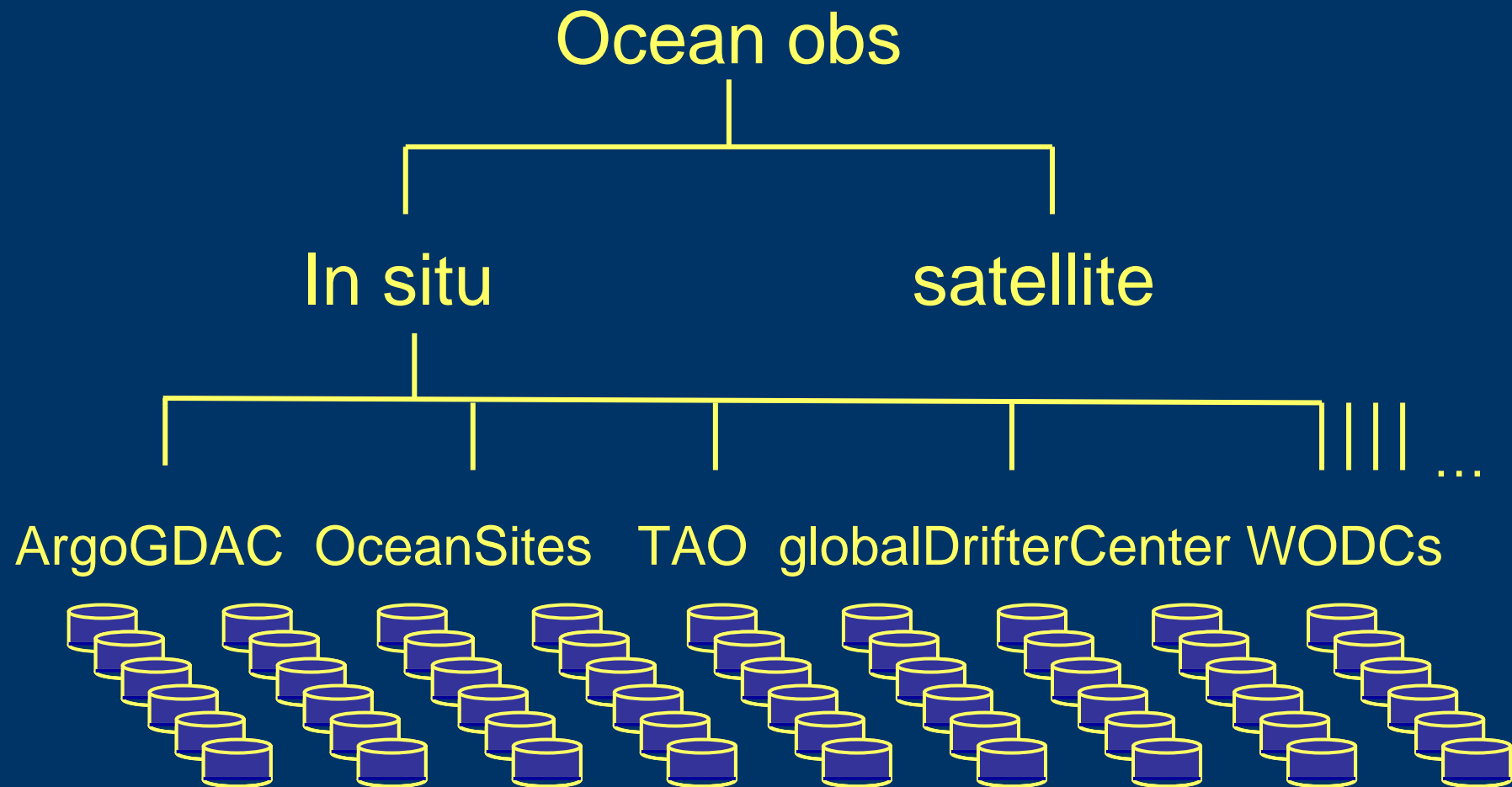
A challenge and an opportunity

First the bad news …

- The scalability of individual file access to 10-100M is "problematic" (to say the least)

Then the good news …

- a single point of entry to a vast collection
- self-describing files
- uniform access techniques for the files
- many clients / reference libraries
- granule-level metadata included

How should we regard the scalability challenge?

> "…*technology develops cumulatively, rather than in…heroic acts …*"
>
> Jared Diamond, 1997 (Pulitzer Prize)

Our actions need to be incremental ("Heroic" visions should guide us)

"Don't let perfect be the enemy of good"

("Don't let tomorrow's customers be the enemy of today's")
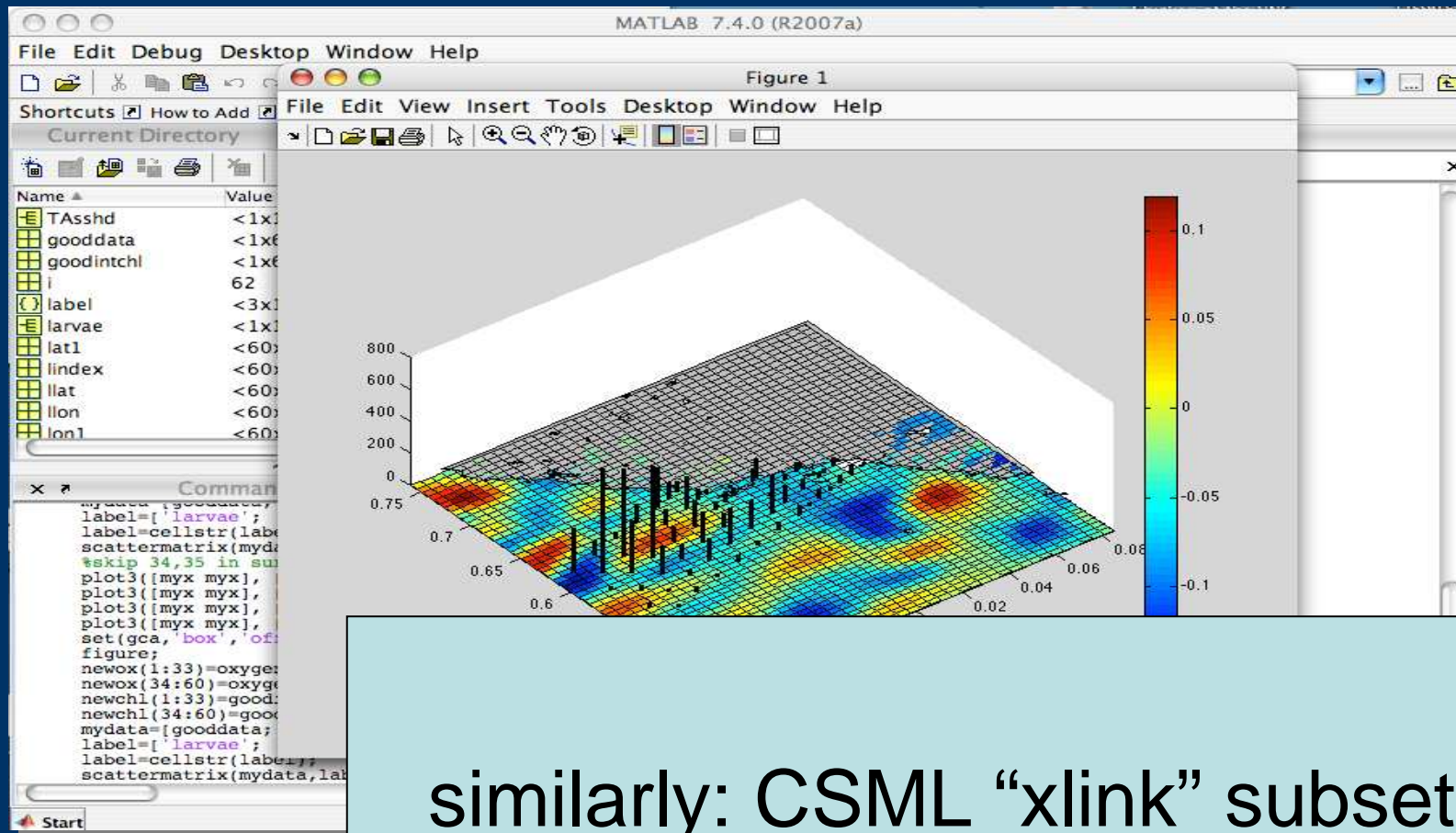
challenge ➜ opportunity

Data Discovery:
- THREDDS crawlers: adapt existing ones from ESG, IMOS, …
- Improve the crawling process with THREDDS advancements
  - Promote selected metadata into the XML
    - Bounding box, parameter names, last mod date
- Multiple catalog/search options exist today

a functioning system for small subsets

Data accessed directly from the supplier …
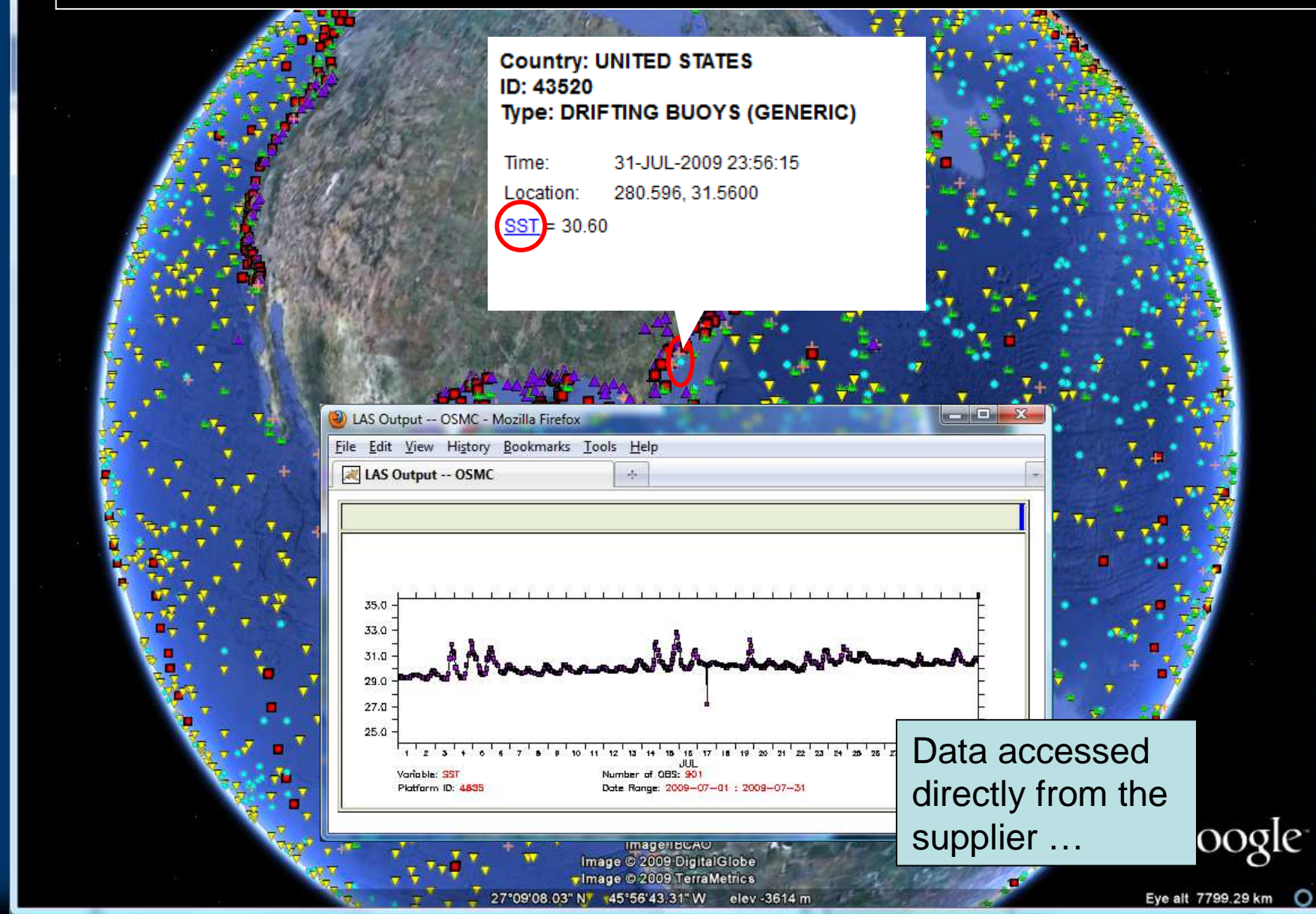
Scriptable clients

Larval densities over Sea Surface Height anomalies
from remote data

similarly: CSML "xlink" subsets,
wget, …

challenge ➔ opportunity

Data access - Possibilities to make it scalable:

- Harvest subsets onto local storage and use
  - DAPPER, ERDDAP, WFS, SOS, (zipped FTP) (what we do today but more general)
    - AND / OR –
- Define "aggregation" for collections of individual obs files into "point data collections" (using by-index style)
    - AND / OR –
- Collaborate with Unidata and OPeNDAP Inc. to make some of these "standard" capabilities in the servers.

challenge ➔ opportunity

– AND / OR –

– Join forces with Unidata and OPeNDAP Inc.
➔ turn some of the preceding into "standard" capabilities in the servers.

challenge ➔ opportunity

– AND / OR –

– Move the data and the processing onto cloud resources

(cost is falling … and already not high when balanced against the value of the data)
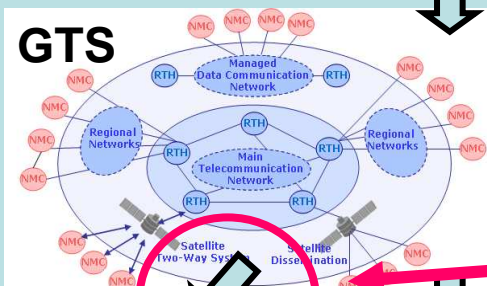
From whence the resources?

Good question.

No easy answer.

A great deal can be accomplished by coordinating work under existing resources

Two additional thoughts

Obs

GTS

Telemetry

Data centers

assimilation centers

archive centers

data assembly centers

GTS access for the common man

 - a ¾ solved problem (lets finish it)

Convergence on uniform file formats: netCDF-CF

The significance of aggregation

"file-at-a-time procedural data analysis is nearing the breaking point"

Jim Grey, (http://research.microsoft.com/pubs/64537/tr-2005-10.pdf)

… yet replacing files by services is still some distance away

- Aggregation through services is an incremental, cumulative approach
- "Files" become "data sets" (virtual files)
- "Physical data independence" is achieved

(a mature solution today for satellites and models)

Discussion

(thank you)

Where funding is needed for integration

narrow   **broad**

Single platform systems   **Integration Components**   **Heroic integration**
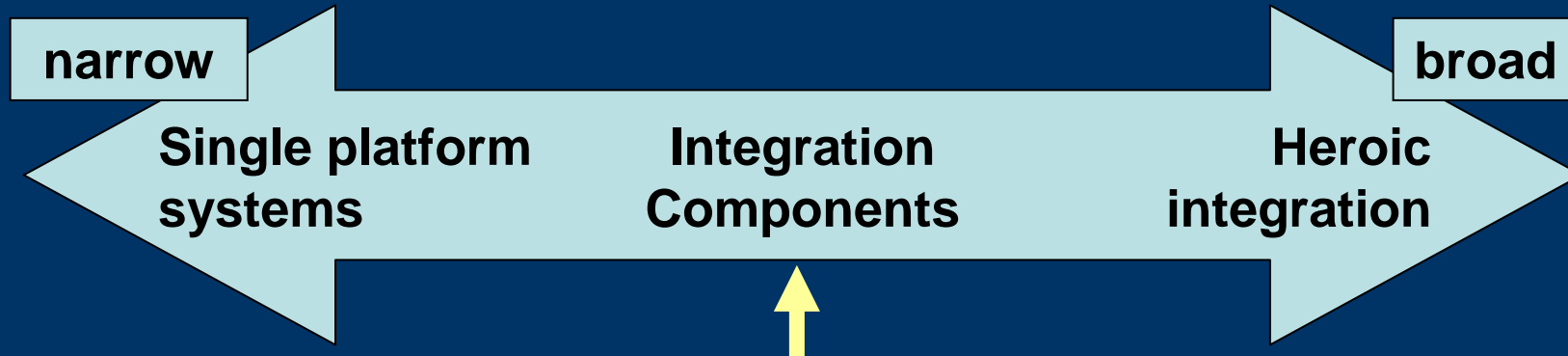
We need a business plan
to help support the
integration components
we depend upon